A French Human Reference Corpus for Multi-Document Summarization and Sentence Compression

Claude de Loupy(1), Marie Guégan(1), Christelle Ayache(1),
Somara Seng(1), Juan-Manuel Torres Moreno(2, 3)

(1) Syllabs
15 rue Jean-Baptiste Berlier, 75013 Paris, France
{loupy, guegan, ayache, seng}@syllabs.com

(2) Laboratoire Informatique d'Avignon (UAPV)
F-84911 Avignon, France
juan-manuel.torres@univ-avignon.fr

(3) Ecole Polytechnique de Montréal
C.P. 6079, succ. Centre-ville, Montréal (Québec) Canada

Abstract

This paper presents two corpora produced within the RPM2 project: a multi-document summarization corpus and a sentence compression corpus. Both corpora are in French. The first one is the only one we know in this language. It contains 20 topics with 20 documents each. A first set of 10 documents per topic is summarized and then the second set is used to produce an update summarization (new information). 4 annotators were involved and produced a total of 160 abstracts. The second corpus contains all the sentences of the first one. 4 annotators were asked to compress the 8432 sentences. This is the biggest corpus of compressed sentences we know, whatever the language. The paper provides some figures in order to compare the different annotators: compression rates, number of tokens per sentence, percentage of tokens kept according to their POS, position of dropped tokens in the sentence compression phase, etc. These figures show important differences from an annotator to the other. Another point is the different strategies of compression used according to the length of the sentence.

1. Introduction

Since the "Road Map for Summarization Research" (Baldwin *et al.*, 2000), several corpora dedicated to the evaluation of multi-document summarization have been compiled: DUC evaluations (from 2001 to 2007) followed by the TAC evaluations (2008-2009) (Dang & Owczarzak, 2008) or some specific developments (Sekine & Nobata, 2003). Automatic multi-document summarization has been explored using various techniques. For example, methods based on the Rhetorical Structure Theory (Mann and Thompson, 1988) consist in deleting the less important textual units by using the discursive structure of a text. These methods (Ono, Sumita and Miike, 1994; Marcu, 1997) use the rhetorical structure trees of a text and compute different node scores. Each node corresponds to a segment of the text. The scores allow determining the segments to select or to cut in order to generate the summary.

There also exist sentence compression-dedicated corpora (Knight & Marcu, 2000; Clarke & Lapata, 2008). Sentence compression consists in removing lexical units that are not important enough in the sentence to change or distort its main meaning. The compressed sentence must be grammatically well-formed and must not be rewritten using new words.

Most of these corpora are in English and we think it is necessary to build such corpora for other languages since the performances of the different methods are likely to change from one language to another.

The RPM2 project[1] is an exploratory industrial research project whose purpose is the development of automatic methods for multi-document summarization including text, audio and video. In order to evaluate our work on textual compression, we developed our own corpora for French[2].

The multi-document summarization corpus contains 20 topics with 20 documents each. The summaries were built as a two-step process: first, as a classic multi-document summarization on an initial set of 10 documents per topic, then as an *update* summarization process given the first set of documents on a second set of 10 documents per topic.

The sentence compression corpus was built using the 400 documents of the first one. It contains 8432 sentences, with 4 annotators per sentence. As far as we know, this is significantly more than any other sentence compression corpus, whatever the language. A large sentence corpus both allows a true learning procedure and a reliable evaluation phase.

The second section of this paper presents the creation of the multi-document summarization corpus. A few statistics on this corpus are given in section 2.2. The third section deals with the making of the sentence compression corpus. Statistics for this corpus are given in section 3.3.

[1] This research was supported by the ANR (French National Research Agency), ANR RPM2 project grant number: ANR-07-AM-008.

[2] The corpus is available at http://lia.univ-avignon.fr/rpm2

The last section draws some conclusions and mentions further works.

2. The Multi-Document Summarization Corpus

2.1. Corpus Compilation

On the whole, the protocol used for this corpus follows the protocol of TAC 2008. 20 topics related to journalistic events were manually chosen. Each topic is associated with 2 sets of 10 documents each. The first set is used for classic summarization processes, whereas the second set is used for an incremental *update* summarization, knowing the first set.

The documents were extracted from several French newspapers. The corpus contains 400 French news articles from January to September 2009. It is available in three formats: text, XML and HTML. Table 1 shows the list of topics.

01	Ingrid Bétancourt	*French journalist held hostage*
02	Caisse d'Epargne	*French bank*
03	Crise bancaire	Banking crisis
04	Dalaï Lama	Dalai Lama
05	Fichier Edvige	*A government database*
06	JO de Pékin	Beijing Olympic Games
07	Jérôme Kerviel	*French trader*
08	Lance Armstrong	Lance Armstrong
09	La loi Leonetti	*French Law*
10	Le petit Mohamed	*Abandoned child*
11	Obama président	Obama for president
12	Licenciement de Patrick Poivre d'Arvor	*A French TV journalist's dismissal*
13	Le temple de Preah Vihear	Preah Vihear Temple
14	Election au PS	Socialist party elections
15	Grossesse Rachida Dati	*French minister's pregnancy*
16	Rachida Dati et les magistrats	*Minister Dati and the magistrates*
17	Réforme du lycée	High school reform
18	Réforme de l'audiovisuel public	Public broadcasting reform
19	Relance de l'économie	Pump priming
20	Crise au Tibet	Crisis in Tibet

Table 1: The 20 topics of the corpus

Four annotators were chosen randomly among French natives. They were asked to produce a summary (abstract) of the first set of ten documents for each topic. These summaries contain from 90 to 100 words each[3]. A second summary was then produced for the second set of ten documents. For the latter, annotators were asked to provide new information only, given that he/she already knew the first set of documents. This task is similar to the TAC 2008 update summarization task: "*produce short (~100 words) multi-document update summaries of Newswire articles under the assumption that the user has already read a set of earlier articles. The purpose of each update summary will be to inform the reader of new information about a particular topic.*"[4]

2.2. Quantitative Study of the Corpus

The summarization corpus contains 400 documents and 160 summaries: 20 topics, 2 summaries per topic, and 4 annotators.

Table 2 gives a few figures describing the original corpus and the summaries. Each summary corresponded to a single set of 10 documents. Therefore figures for the original corpus were computed set by set and averaged over the sets to remain consistent with the summaries. In this table, words represent tokens distinct from punctuation. The average length is 104 words per summary according to our automatic tokenization[3].

	Original	**A1**	**A2**	**A3**	**A4**
Mean nb sent.	205	5.3	5.5	4.3	5.9
Mean nb tokens	5640	114	113	115	114
Mean nb words	4882	104	104	104	102
Mean nb tokens per sentence	28.7 ±6.8	22.4 ±6.8	22.0 ±6.7	27.6 ±7.7	19.7 ±9.0

Table 2: General figures for the summarization corpus

The mean number of tokens per sentence is much lower in the summaries than in the original corpus, where it corresponds to the usual length for a newspaper sentence in French. As we will see in section 3.3, it is very close to what can be found in manually compressed sentences. The only exception is annotator 3, who preferred to use fewer but longer sentences.

POS	**Original**	**A1**	**A2**	**A3**	**A4**
noun	**19.2**	19.8	20.8	20.6	19.7
preposition	**15.9**	16.6	16.9	18.3	15.8
punctuation	**14.4**	8.4	7.6	9.7	9.8
verb	**13.6**	16.5	16.1	13.7	15.0
determiner	**11.3**	12.1	13.6	11.6	12.2
pronoun	**5.8**	5.1	4.0	5.6	5.0
adjective	**5.4**	4.8	5.6	5.6	6.2
adverb	**4.2**	4.6	2.8	4.9	4.3
conjunction	**3.1**	3.3	4.4	3.4	4.2
number	**1.8**	1.9	1.4	1.4	1.3

Table 3: Proportion of POS tags in the original corpus and in the summaries (%)

[3] Note that, to be consistent with related work, a word is here defined as a sequence of characters between spaces. "*Le chat d'Antoine*" is therefore counted as 3 words instead of 4 as it should be in a French parser. This explains that figures in section 2.2 are above 100.

[4] http://duc.nist.gov/duc2007/tasks.html

Table 3 shows the proportion of part-of-speech tags (POS) found in the original corpus as well as in the summaries produced by the annotators. These figures give an insight of the use of grammatical categories in summaries compared to news articles. The corpus was tagged using TreeTagger (Schmid, 1994).

We observe some differences: the summaries contain more verbs, determiners, prepositions and nouns as well as less punctuation, adverbs and pronouns than the original corpus. As we will see in section 3.3, these figures are interestingly similar to those obtained for the sentence compression corpus.

3. The Sentence Compression Corpus

In this work, sentences are compressed only by removing words. No other change is allowed.

3.1. Similar Corpora

The Ziff-Davis Corpus (Knight & Marcu, 2000) contains 1067 sentence pairs extracted from computer product-related articles. The Written News Compression Corpus[5] (Clarke & Lapata, 2006) is a collection of 1629 compressed sentences from 82 articles of the British National Corpus (BNC), the LA Times, and the Washington Post.

Two French corpora already exist: Myriam and UNIVERSITE. The Myriam corpus was compiled by Michel Gagnon, then adapted and used by (Waszack & Torres-Moreno, 2008) who needed a corpus to train their system in French. It contains 219 human-compressed sentences extracted from narrative texts. The UNIVERSITE corpus (Torres-Moreno, 2010) is composed of 500 heterogeneous original-compressed sentence pairs extracted from blogs, e-mails, tracts and Wikipedia documents.

3.2. Corpus Creation

For the RPM2 sentence compression corpus, we used the 400 documents of the summarization corpus. It represents 8432 sentences. Each of the four annotators compressed all the sentences by hand. Since certain annotators had already produced the summaries, they knew the texts and could be influenced by their previous work, so a prior shuffling of the sentences was conducted.

The annotators were asked to compress the sentences by removing the elements that could be considered less relevant to the main meaning. Hence, they had to respect the following points:

- *Grammaticality*: annotators had to ensure that the compressed sentence was grammatically well-formed.
- *Importance*: annotators had to ensure that the main meaning of the original sentence remained unchanged and that the compressed sentence was still coherent.

These criteria are widely used in the sentence compression community for manually evaluating results.

Examples of compressed sentences can be found in figure 1. The first sentence is the original one. The following sentences correspond to its four compressions. We notice that the annotators adopted different word-deletion strategies. For instance, annotator 4 was far more aggressive than annotator 1, who deleted fewer words.

Les banques françaises n'ont pas publié de chiffres précis sur leur exposition à Lehman Brothers mais ont diffusé des messages au marché laissant entendre clairement que celle-ci était limitée et bénéficiait, pour ce qui est du risque de contrepartie sur des transactions de marché, de sûretés sous forme de collatéral.

A1 *Les banques françaises n'ont pas publié de chiffres sur leur exposition à Lehman Brothers mais ont diffusé des messages laissant entendre que celle-ci était limitée et bénéficiait de sûretés.*

A2 *Les banques françaises n'ont pas publié de chiffres sur leur exposition mais ont diffusé des messages laissant entendre que celle-ci était limitée et bénéficiait de sûretés sous forme de collatéral.*

A3 *Les banques françaises n'ont pas publié de chiffres sur leur exposition à Lehman Brothers mais ont diffusé des messages que celle-ci était limitée.*

A4 *Les banques n'ont pas publié de chiffres sur leur exposition à Lehman Brothers.*

Figure 1: Examples of compressed sentences extracted from the RPM2 sentence compression corpus

3.3. Quantitative Study and Agreement

Our corpus contains 8432 sentences and about 26.2 tokens per uncompressed sentence.

The *compression rate* (CR) of a textual unit (corpus or sentence) is defined as the overall percentage of words *kept* in the compressed version of this unit. Uncompressed units therefore have a CR of 100%. Table 4 shows high overall compression rates for our sentence compression corpus, between 71% and 84% compared to 56% and 70% in existing written corpora (Cohn & Lapata, 2009). The compression rate reaches 86% for annotation 2 when computed at the sentence level and averaged over all sentences. Annotators did not compress the sentences very much compared to previous work.

Annotator	A1	A2	A3	A4
Compression rate (%)	75.0	84.4	71.0	79.0
Tokens per sentence after compression	19.7	22.1	18.6	20.7
Mean CR per sentence	80.3	86.5	77.1	82.5
Smallest CR (%) reached for a sentence	12.0	26.9	7.8	17.1
Untouched sentences	2119	2142	1630	1775

Table 4: General figures for the sentence compression corpus

[5] http://homepages.inf.ed.ac.uk/s0460084/data

Figure 2 shows the distribution of compression rates for all annotators. A great deal of sentences remained unchanged, about a quarter of them, as depicted in table 4 (*Untouched sentences*). The compression rate decreases relatively softly to 7.8% for annotator 3.

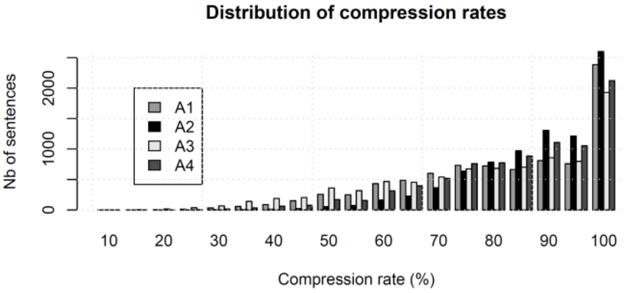

Figure 2: Distribution of compression rates

In the distribution of the size of manually deleted segments (see figure 3), we observe an exponential drop, characterized by a large number of single-word deletions. In the corpus, large deletion spans (> 40 words) appeared mostly in the case of enumerations. Annotator 2 deleted much smaller segments, leading to a higher compression rate (see table 4). This annotator was very shy in all respects: highest number of uncompressed sentences, highest mean compression rate per sentence, and highest minimum compression rate on the corpus (a quarter of the sentence).

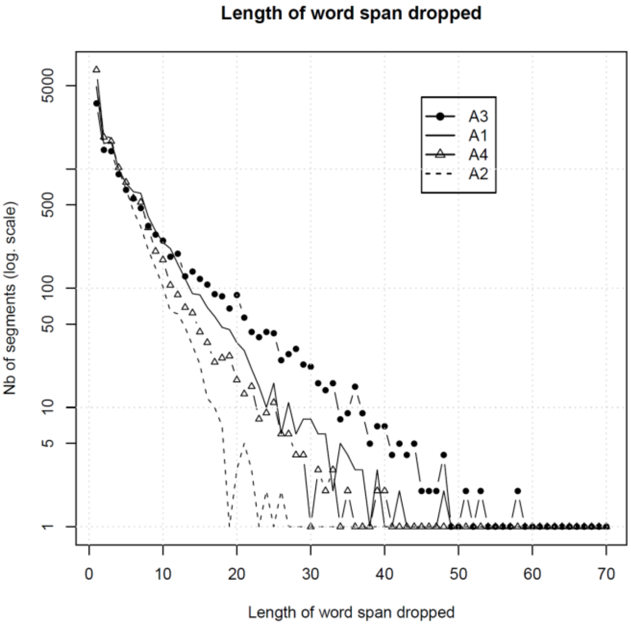

Figure 3: Length of word span dropped (log scale)

Annotator 3 was much more active at dropping words, followed by annotators 1 and 4. Interestingly, figure 3 is quite different from its equivalent in (Clarke & Lapata, 2006), showing a much softer decrease and only rare deletions of more than 10 consecutive words. Their corpus, Broadcast News, is a set of manually transcribed broadcast news stories. It contains 1370 sentences, with about 19 words per sentence. Three annotators compressed the sentences down to 73% in average, which was similar to our compression rates. Since sentences were longer in our case, we expected the word spans to be greater too.

Tables 5 and 6 show the correlation and contingency tables between annotators. The correlation was computed using the *phi coefficient* for two binary variables, in this case equivalent to the Pearson correlation coefficient. Figures were computed considering (a) all tokens and (b) only tokens that were dropped by at least 1 annotator. Correlation over all tokens is low, especially between annotators 2 and 3. Annotators 1 and 4 show the strongest correlation at 0.53. The correlation is especially low when we consider dropped tokens only. Annotator 3, who deleted most tokens, is almost independent from the others. In table 3, we observe that annotators dropped 25% (A1), 15% (A2), 29% (A3) and 21% (A4) of the words. They altogether agreed on 62% of the tokens, 56 % of which were kept and 6% dropped.

Annotators	**1 & 2**	**1 & 3**	**1 & 4**	**2 & 3**	**2 & 4**	**3 & 4**
Phi (a)	0.43	0.43	0.53	0.34	0.44	0.38
Phi (b)	0.16	-0.07	0.24	-0.02	0.21	-0.08

Table 5: Correlation between annotators (Phi coefficient) (a) over all tokens (b) over dropped tokens

A1-A2	**Drop**	**Keep**	**Sum**	**A2-A3**	**Drop**	**Keep**	**Sum**
Drop	10,7	14,3	24,9	**Drop**	10,1	5,5	15,5
Keep	4,9	70,2	75,1	**Keep**	18,8	65,6	84,5
Sum	15,5	84,5	100	**Sum**	28,9	71,1	100
A1-A3	**Drop**	**Keep**	**Sum**	**A2-A4**	**Drop**	**Keep**	**Sum**
Drop	15,7	9,2	24,9	**Drop**	9,7	5,8	15,5
Keep	13,2	61,9	75,1	**Keep**	11,3	73,2	84,5
Sum	28,9	71,1	100	**Sum**	21,0	79,0	100
A1-A4	**Drop**	**Keep**	**Sum**	**A3-A4**	**Drop**	**Keep**	**Sum**
Drop	14,5	10,4	24,9	**Drop**	13,0	15,9	28,9
Keep	6,4	68,6	75,1	**Keep**	7,9	63,1	71,1
Sum	21,0	79,0	100	**Sum**	21,0	79,0	100

Table 6: Contingency table between annotators (%)

3.4. Linguistic Study

We eventually observed the nature and position of dropped elements in the sentence.

A number of n-grams were systematically deleted by some or all annotators, such as: forenames (*Angela*), adverbs (*clairement*), adverbial phrases (*par ailleurs*), temporal phrases (*le 6 octobre*, *vendredi soir*, *âgé de*, *en fin d'après-midi*), phrases preceding names (*l'ancien président Nicolas*), citations blanks (*[…]*). Phrases between parentheses were also systematically dropped, except for cases where it appeared inside a quote.

Table 7 shows compression rates for different POS, averaging over all annotators. Sentence tags (“?”, “!”, “.”) and double quotes were almost always kept. About a third of adjectives, punctuation marks (putting aside double quotes and sentence tags), numbers, and adverbs were deleted. In contrast, the percentage of verbs and determiners significantly increased with the compression.

POS	**% in corpus**		**CR for this POS**
	Before	**After**	
verb	13,4	15,1	87,2
determiner	11,2	11,8	81,5
pronoun	6,0	6,2	80,0
preposition	15,8	15,7	76,8
noun	19,1	18,9	76,5
conjunction	3,1	3,0	74,8
adjective	5,3	4,8	70,8
punctuation	10,0	8,9	69,5
number	1,8	1,4	63,8
adverb	4,2	3,0	55,5

Table 7: Compression rate per POS

Our last figure represents the location of dropped words inside sentences (figure 4). We made a distinction between short (less than 17 tokens), medium-size (between 17 and 31 tokens) and long sentences (more than 31 tokens). Following this definition, the corpus contains approximately the same number of sentences in each category (2804, 2713 and 2915). The horizontal axis represents the relative position of tokens in sentences, where 0 and 100 respectively denote the beginning and ending of the sentence.

First, we notice that all annotators dropped approximately the same number of tokens at the very beginning of the sentence, no matter the size of the sentence. Indeed, news article sentences often begin with introductory words specifying time (“*Demain*, ...”) or coordinating it to the preceding sentence in the original text (“*Cependant*...”). In this task, sentences had been made independent from each other with all context removed, which may explain why annotators dropped this type of information.

Then, we notice that deletion profiles differ according to the size of the sentence.

The annotators show very similar profiles for short sentences (fig. 4.a.), dropping words preferably at the beginning of the sentence, uniformly inside the sentence, and systematically keeping the end of the sentence. This can easily be explained by the fact that the end token is the irremovable sentence tag. The same observation can be made in figures 4.b. and 4.c.

Annotator 2 is the only one who seems to have stuck to this strategy for medium-size sentences. The others chose to delete the end of sentences more often.

In the case of long sentences, annotators 1, 2 and 4 show very similar profiles: deletions occurred uniformly across the sentence with a small increase towards the end. In contrast, annotator 3 deleted many more words in the second half of the sentence. Surprisingly, this annotator has almost the same profile as annotator 1 in the first third of the sentences, dropping even less tokens. His drastic compression technique thus seems to rely on deleting the last part of sentences. This annotator took advantage of the fact that news articles often state the relevant information in the first part of the sentence, giving more details at the end.

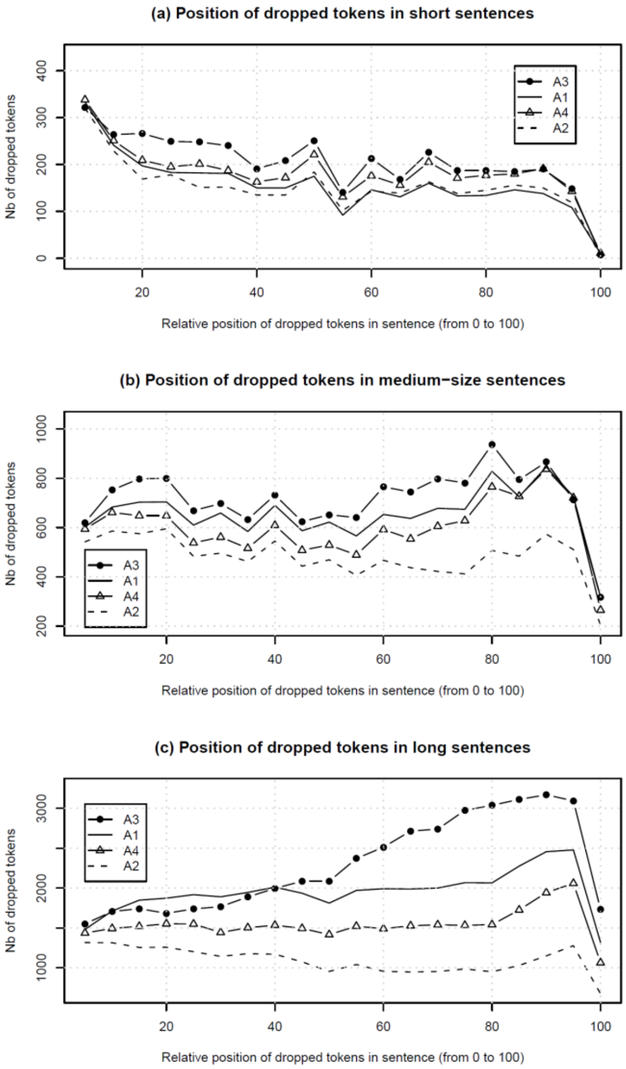

Figure 4: Position of dropped tokens in the sentence according to its size: (a) short (b) medium (c) long

4. Conclusion

The first corpus concerns multi-document summarization and is highly comparable with the TAC 2008 corpus. As far as we know, it is the first French corpus of this type.

The sentence compression is not the first French corpus of this type but it is larger than any other similar corpus.

The first corpus was used for a summarization evaluation (Boudin & Torres-Moreno, 2009) and we hope it will be widely used. In particular, we think it is very important to compare the performances of the different methods, across languages.

5. Acknowledgements

This work has been done in collaboration with Sinequa and a special thank goes to Frederik Cailliau for his support and advice. Juan-Manuel Torres specially thanks Iria da Cunha for the discussion and references over the corpus construction.

6. References

Baldwin, B., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., et al. (2000). An Evaluation Road Map for Summarization Research. doi: 10.1.1.16.4807.

Boudin, F., Torres-Moreno, J. (2009). Résumé automatique multi-document et indépendance de la langue : une première évaluation en français. In *Actes de TALN'2009* (Vol. c). Senlis, France.

Clarke, J., Lapata, M. (2006). Models for Sentence Compression: A Comparison across Domains, Training Requirements and Evaluation Measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 377--384.

Clarke, J., Lapata, M. (2008). Global Inference for Sentence Compression: An Integer Linear Programming Approach. In *Journal of Artificial Intelligence Research, 31*, pp. 399--429.

Cohn, T., Lapata, M. (2009). Sentence Compression as Tree Transduction. *Journal of Artificial Intelligence Research*, Vol. 34, pp. 637--674.

Dang, H. T., Owczarzak, K. (2008). Overview of the TAC 2008 Update Summarization Task. In *Proceedings of Text Analysis Conference*, pp. 1--16.

Knight, K., Marcu, D. (2000). Statistics-Based Summarization - Step One: Sentence Compression. In *Proceedings of the 7th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pp. 703--710. AAAI Press / The MIT Press.

Mann, C., Thompson S. (1988) Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3), pp. 243--281

Marcu, D. (1997) From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97. Workshop on Intelligent Scalable Text summarization*, pp. 82--88. Madrid, Spain

Ono, K., Sumita, and Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In *Proceedings of the International Conference on Computational Linguistics* (Coling-94), pp. 344--348

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pp. 44--49. Manchester, UK.

Sekine, S., Nobata, C. (2003). A survey for multi-document summarization. In *Human Language Technology Conference*.

Torres-Moreno J-M. (2010). The « *Université* » human sentences compressed corpus. Technical Report RT-LIA-2010.03.23, Laboratoire Informatique d'Avignon (UAPV), March.

Waszack, T., Torres-Moreno J-M. (2008). Compression entropique de phrases contrôlée par un perceptron. In *Proceedings of JADT 2008*, pp. 1163--1173.